

QUT Digital Repository:  
<http://eprints.qut.edu.au/>



Yu, Zuguo and Zhou, Li-Qian and Anh, Vo V. and Chu, K. H. and Li, C. P. and Chen, Y. J. (2007) Distance-based analysis to reveal vertebrate phylogeny without sequence alignment using complete mitochondrial genomes. In Callaos, N. and Lesso, W. and Zinn, C. and Zmazek, B., Eds. *Proceedings 11th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2007*, pages pp. 206-211, Florida, USA.

© Copyright 2007 (please consult author)

# Distance-based analyses to reveal vertebrate phylogeny without sequence alignment using complete mitochondrial genomes

**Z.G. Yu**

School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434,  
Brisbane, Queensland 4001, Australia. Email: [z.yu@qut.edu.au](mailto:z.yu@qut.edu.au)  
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.

**L.Q. Zhou**

School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China.

**V.V. Anh**

School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434,  
Brisbane, Queensland 4001, Australia. Email: [v.anh@qut.edu.au](mailto:v.anh@qut.edu.au)

**K.H. Chu and C.P. Li**

Department of Biology, Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China.

**Y.J. Chen**

Department of Applied Computer Science, University of Winnipeg, 515 Portage Ave., Winnipeg,  
Manitoba R3B 2E9, Canada.

## Abstract

There have been a number of recent attempts to develop methodologies that do not require sequence alignment for deriving species phylogeny based on overall similarities of the complete genomes. The mitochondrial genomes have provided much information on the evolution of this organelle and have been used for phylogenetic reconstruction by various methods with or without sequence alignment. In this paper we introduce three fast algorithms, namely, dynamical language model with correlation distance, Fourier transform with Kullback-Leibler divergence distance, log-correlation distance, for deriving vertebrate phylogeny based on mitochondrial genomes. The distance-based analyses show that the mitochondrial genomes are separated into three major clusters corresponding to mammals, fish, and Archosauria (including birds and reptiles) respectively. The interrelationships among the mitochondrial genomes are roughly in agreement with the current understanding on the phylogeny of vertebrates revealed by the traditional approaches.

**Keywords:** vertebrate phylogeny; mitochondrial genome; dynamical language model; Fourier transform; correlation distance; Kullback-Leibler divergence distance

## 1. Introduction

Traditional molecular phylogenies are often based on sequences from only one or a few genes (e.g. Woese *et al.* 1990). It is generally accepted that whole genome sequences are better tools for studying evolution (Eisen and Fraser 2003). There have been a number of attempts recently to develop methodologies without sequence alignment for deriving species phylogeny based on overall similarities of complete genomes. These methods include information-based analysis (Li *et al.* 2001; Yu and Jiang 2001), principal component analysis (Edwards *et al.* 2002), singular value decomposition (SVD) (Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002), Markov model (Qi, Wang and Hao 2004; Qi, Luo and Hao 2004), fractal analysis (Yu, Anh and Lau 2003, 2004; Yu *et al.* 2003) and dynamical language model (Yu *et al.* 2005).

The phylogenetic signal in the protein sequences is often obscured by noise and bias (Charlebois, Beiko and Ragan 2003). There is always some randomness in the composition of protein sequences, revealed by their statistical properties at single amino acid or oligopeptide level (Weiss, Jimenez and Herzel 2000). By overcoming the problem of noise and bias in the protein sequences through the use of suitable models, whole-genome trees have now largely converged to the rRNA-sequence tree (Charlebois, Beiko and Ragan 2003). Simple correlation analyses of complete genome sequences using Markov model (Qi, Wang and Hao 2004) and dynamical language model (Yu *et al.* 2005) without sequence alignment have been developed. The analyses based on these two methods

using 103 prokaryotes and 6 eukaryotes have yielded trees separating the three domains of life, Archaea, Eubacteria and Eukarya, with the relationships among the taxa agreeing with those based on traditional analyses (Qi, Wang and Hao 2004, Yu *et al.* 2005). These two methods were then used to analyze the phylogenetic relationships of complete chloroplast genomes (Chu *et al.* 2004; Yu *et al.* 2005). A simplified method from that in Qi, Wang and Hao (2004) was used to analyze rRNA gene sequences as molecular barcodes (Chu, Li and Qi 2006).

Mitochondrial genes and genomes have long been a major focus in molecular evolution, and these genomes are excellent candidates for demonstrating the power of evolutionary genomics. Mitochondrial DNA has proven to be a useful tool for phylogenetic reconstruction, especially when complete genomes are considered (Reyes, Pesole and Saccone 1998). A correlation analysis based on a different transformation of compositional vectors to reveal phylogeny using vertebrate mitochondrial genomes was also reported (Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002). In the present paper, we introduce some distance methods including the dynamical language model approach (Yu *et al.* 2005) for vertebrate phylogenetic analysis using a large number of mitochondrial genomes.

## 2. Materials

**Genome Data Set:** In order to explore the feasibility of our methods, we use the 64 complete mitochondrial genomes data set used by Stuart, Moffet and Leader (2002). The whole DNA sequences (including protein-coding and non-coding regions), all protein-coding DNA sequences and all protein sequences of these complete genomes were obtained from the NCBI genome database (<http://www.ncbi.nlm.nih.gov/genbank/genomes>). Species represented in the analysis include the following: *Alligator mississippiensis* (Amis), *Artibeus jamaicensis* (Ajam), *Aythya Americana* (Aame), *Balaenoptera musculus* (Bmus), *Balaenoptera physalus* (Bphy), *Bos taurus* (Btau), *Canis familiaris* (Cfam), *Carassius auratus* (Caur), *Cavia porcellus* (Cpor), *Ceratotherium simum* (Csim), *Chelonia mydas* (Cmyd), *Chrysemys picta* (Cpic), *Ciconia boyciana* (Cboy), *Ciconia ciconia* (Ccic), *Corvus frugilegus* (Cfru), *Crossostoma lacustre* (Clac), *Cyprinus carpio* (Ccar), *Danio rerio* (Drer), *Dasyatis novemcinctus* (Dnov), *Didelphis virginiana* (Dvir), *Dinodon semicarinatus* (Dsem), *Equus asinus* (Easi), *Equus caballus* (Ecab), *Erinaceus europaeus* (Eur), *Eumeces egregius* (Eegr), *Falco peregrinus* (Fper), *Felis catus* (Fcat), *Gadus morhua* (Gmor), *Gallus gallus* (Ggal), *Gorilla gorilla* (Ggor), *Halichoerus grypus* (Hgr), *Hippopotamus amphibius* (Hamp), *Homo sapiens* (Hsap), *Latimeria chalumnae* (Lcha), *Loxodonta africana* (Lafr), *Macropus robustus* (Mrob), *Mus musculus* (Mmus), *Mustelus manazo* (Mman), *Myoxus glis* (Mgli), *Oncorhynchus mykiss* (Omyk), *Ornithorhynchus anatinus* (Oana),

*Orycteropus afer* (Oafe), *Oryctolagus cuniculus* (Ocu), *Ovis aries* (Oari), *Paralichthys olivaceus* (Poli), *Pelomedusa subrufa* (Psub), *Phoca vitulina* (Pvit), *Polypterus ornatipinnis* (Porn), *Pongo pygmaeus abelii* (Ppyg), *Protopterus dolloi* (Pdol), *Raja radiata* (Rrad), *Rattus norvegicus* (Rnor), *Rhea americana* (Rame), *Rhinoceros unicornis* (Runi), *Salmo salar* (Ssal), *Salvelinus alpinus* (Salp), *Salvelinus fontinalis* (Sfon), *Scyliorhinus canicula* (Scan), *Smithornis sharpei* (Ssha), *Squalus acanthias* (Saca), *Struthio camelus* (Scam), *Sus scrofa* (Sscr), *Talpa europaea* (Teur), and *Vidua chalybeata* (Vcha). The words in the brackets are the abbreviations of the names of these organisms used in our phylogenetic trees (Figs. 1, 2 and 3).

## 3. Methods

In this paper, three kinds of data from complete genomes are analysed. They are the whole DNA sequences (including protein-coding and non-coding regions), all protein-coding DNA sequences and the amino acid sequences of all protein-coding genes.

We regard DNA sequences of the 4 nucleotides and protein sequences of the 20 amino acids as symbolic sequences. Then we consider strings with fixed length  $K$ , called  $K$ -strings. There are a total of  $N = 4^K$  (for DNA sequences) or  $20^K$  (for protein sequences) possible types of  $K$ -strings. Assume the length of a DNA or protein sequence is  $L$ . We use a window of length  $K$  and slide it through the sequences by shifting one position at a time to determine the frequencies of each of the  $N$  types of  $K$ -strings in this sequence. The observed frequency  $p(s_1s_2...s_K)$  of a  $K$ -string  $s_1s_2...s_K$  is defined as  $p(s_1s_2...s_K) = n(s_1s_2...s_K) / (L - K + 1)$ , where  $n(s_1s_2...s_K)$  is the number of times that  $s_1s_2...s_K$  appears in this sequence. For the DNA or amino acid sequences of the protein-coding genes, denoting by  $m$  the number of protein-coding DNA sequences or the corresponding protein sequences from each complete genome, the observed frequency of a  $K$ -string  $s_1s_2...s_K$  is defined as  $(\sum_{j=1}^m n_j(s_1s_2...s_K)) / (\sum_{j=1}^m (L_j - K + 1))$ ; here  $n_j(s_1s_2...s_K)$  means the number of times that  $s_1s_2...s_K$  appears in the  $j$ th protein-coding DNA sequence or protein sequence and  $L_j$  the length of the  $j$ th protein-coding DNA sequence or protein sequence in this complete genome. For all possible  $K$ -strings  $s_1s_2...s_K$ , we use  $p(s_1s_2...s_K)$  as components to form a *composition vector* for a genome. To further simplify the notation, we use  $p_i$  for the  $i$ -th component corresponding to the string type  $i$ ,  $i = 1, \dots, N$  (the  $N$  strings are arranged in a fixed order as the alphabetical order). Hence we construct a composition

vector  $p = (p_1, p_2, \dots, p_N)$  for a genome. Now we introduce three methods to define the distance between two genomes:

**Method 1:** (Dynamical language model with correlation distance). In this method, we consider an idea from the theory of dynamical language that a  $K$ -string  $s_1 s_2 \dots s_K$  is possibly constructed by adding a letter  $s_K$  to the end of the  $(K-1)$ -string  $s_1 s_2 \dots s_{K-1}$  or a letter  $s_1$  to the beginning of the  $(K-1)$ -string  $s_2 s_3 \dots s_K$ . Supposing that we have performed direct counting for all strings of length  $(K-1)$  and the 20 kinds of letters, the expected frequency of appearance of  $K$ -strings is predicted by

$$q(s_1 s_2 \dots s_K) = \frac{p(s_1 s_2 \dots s_{K-1})p(s_K) + p(s_1)p(s_2 s_3 \dots s_K)}{2} \quad (1)$$

where  $q$  denotes the predicted frequency, and  $p(s_1)$  and  $p(s_K)$  are frequencies of nucleotides or amino acids  $s_1$  and  $s_K$  appearing in this genome. Then  $q(s_1 s_2 \dots s_K)$  of all  $4^K$  or  $20^K$  kinds of  $K$ -strings is viewed as the noise background. We then subtract this noise background before performing a cross-correlation analysis through defining

$$X(s_1 s_2 \dots s_K) = \begin{cases} p(s_1 s_2 \dots s_K)/q(s_1 s_2 \dots s_K) - 1, & \text{if } q(s_1 s_2 \dots s_K) \neq 0 \\ 0 & \text{if } q(s_1 s_2 \dots s_K) = 0 \end{cases}$$

The transformation  $X = (p/q) - 1$  has the desired effect of subtraction of random background in  $p$  and rendering it a stationary time series suitable for subsequent cross-correlation analysis. For all possible  $K$ -strings  $s_1 s_2 \dots s_K$ , we use  $X(s_1 s_2 \dots s_K)$  as components to form a composition vector for a genome. To further simplify the notation, we use  $X_j$  for the  $j$ -th component corresponding to the string type  $j$ ,  $j = 1, \dots, N$  (the  $N$  strings are arranged in a fixed alphabetical order). Hence we construct a composition vector  $X = (X_1, X_2, \dots, X_N)$  for genome  $X$ , and likewise  $Y = (Y_1, Y_2, \dots, Y_N)$  for genome  $Y$ . If we view the  $N$  components in vectors  $X$  and  $Y$  as samples of two random variables respectively, the sample correlation  $C(X, Y)$  between any two genomes  $X$  and  $Y$  is defined in the usual way. The distance  $D(X, Y)$  between the two genomes is then defined by  $D(X, Y) = (1 - C(X, Y)) / 2$ .

**Method 2:** (Fourier transform with Kullback-Leibler divergence distance). Fourier transform is widely used to subtract random background in the field of signal analysis. Once we have obtained the composition vector  $p = (p_1, p_2, \dots, p_N)$ , we define the discrete Fourier transform by

$$DFT(f) = \frac{1}{N} \sum_{j=0}^{N-1} p_j e^{-2\pi i f j / N},$$

$f = 0, 1, \dots, N-1$ , and  $i$  is the complex number  $i^2 = -1$ . Then we define

$$X_j = |DFT(j+1)|, \quad j = 1, 2, \dots, N.$$

We use the  $N$ -point fast Fourier transform to get  $X_j$ ,

$j = 1, 2, \dots, N$ . Likewise  $Y = (Y_1, Y_2, \dots, Y_N)$  for genome  $Y$ . If we view the  $N$  components in the vectors  $X$  and  $Y$  as samples of two random variables respectively, the sample Kullback-Leibler divergence  $KL(X|Y)$  between any two genomes  $X$  and  $Y$  is defined as

$$KL(X|Y) = \sum_{i=1}^N X_i \log(X_i / Y_i) \quad \text{when } X_i, Y_i \neq 0.$$

Then we define Kullback-Leibler divergence distance as

$$KLD(X, Y) = \frac{KL(X|Y) + KL(Y|X)}{2}.$$

The Kullback-Leibler divergence (KLD) is an important measure based on information theory (Cover and Thomas 1991).

**Method 3:** (Log-correlation distance). For the composition vectors  $P$  for genome  $X$  and  $Q$  for genome  $Y$ , we define directly the log-correlation distance used in Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002. At first, we define the cosine value of the angle of two vectors  $P$  and  $Q$  as  $\cos \theta = \frac{\langle P, Q \rangle}{\|P\| \times \|Q\|}$ , here  $\langle P, Q \rangle$  means the inner

product of the vectors  $P$  and  $Q$ , and  $\|P\|$  the geometric length of the vector  $P$ . Then we define the distance of the two vectors  $P$  and  $Q$  as

$$d_{pq} = -\log[(1 + \cos \theta) / 2].$$

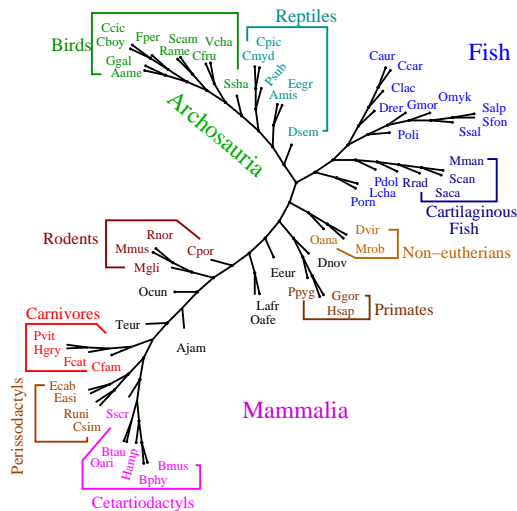
The distance matrices for all the genomes under study using the above three methods are calculated to construct phylogenetic trees. We construct all the trees using the neighbour-joining (NJ) method (Saitou and Nei 1987) in version 3.5c of the PHYLIP package (Felsenstein 1993).

## 4. Results and discussion

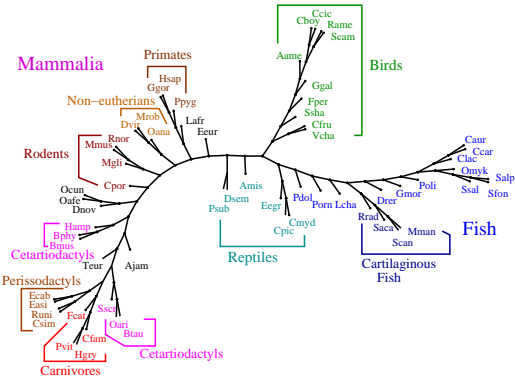
The whole DNA sequences, all protein-coding DNA sequences and all protein sequences from complete mitochondrial genomes of the selected 64 vertebrates were analyzed. The trees of  $K=3$  to 6 based on all protein sequences and the trees of  $K \leq 13$  based on the whole DNA sequences and all protein-coding DNA sequences using the three methods are constructed. After comparing all the trees constructed by the present methods with the traditional classification of the 64 vertebrates (the traditional classification from the KEGG database is available

under “Complete Mitochondrial Genomes” on <http://www.genome.jp/kegg/genes.html>, we find that for the dynamical language model with correlation distance method (method 1), the tree of  $K=11$  based on the whole DNA sequences is the best tree (shown in Fig. 1); for Fourier transform with Kullback-Leibler divergence distance (KLD) approach (method 2), the tree of  $K=5$  using all protein sequences is the best tree (shown in Fig. 2); for the log-correlation distance method (method 3), the tree of  $K=12$  using the whole genome DNA sequences is the best one and we show it in Fig. 3. The distance matrices generated from our analyses are provided upon request via emailing [z.yu@qut.edu.au](mailto:z.yu@qut.edu.au).

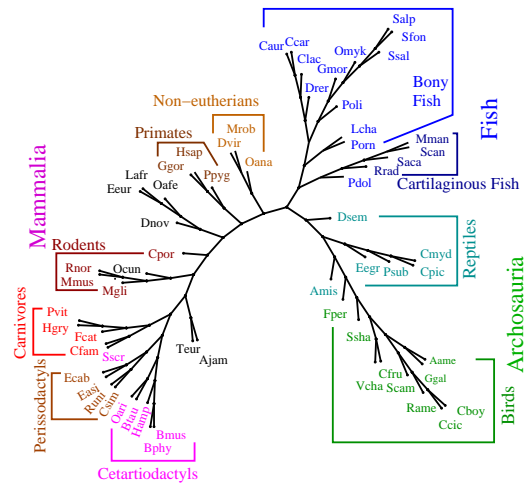
The trees generated are similar in topology to the tree obtained using the SVD method in the case  $K = 4$  (Stuart, Moffet and Leader 2002), and also similar to a recently generated 69 species tree (Pollock *et al.* 2000), placing the vast majority of species into well-accepted groupings. As given in Fig. 1, our distance-based analysis shows that the mitochondrial genomes are separated to three major clusters, corresponding to mammals, fish, and Archosauria (including birds and reptiles), respectively. Within the mammals, cetartiodactyls (cetaceans and artiodactyls), carnivores and perissodactyls (except the elephant (Lafr)) are grouped together as expected (Arnason *et al.* 2000; Murphy *et al.* 2001a, 2001b; Stuart, Moffet and Leader 2002; Xu, Janke and Arnason 1996). These groups form the ferungulates, which together with the mole (Teur) and the bat (Ajam) form a clade as revealed in recent independent analyses (Mouchaty *et al.* 2000; Nikaido *et al.* 2000; Stuart, Moffet and Leader 2002). For the other mammals, non-eutherians



**Fig. 1.** NJ tree of mitochondrial genomes based on the whole DNA sequences using dynamical language model approach in the case  $K=11$ . In this tree birds and reptiles group together as Archosauria.



**Fig. 2.** NJ tree of mitochondrial genomes based on DFT with Kullback-Leibler divergence (KLD) distance in the case  $K=5$  using the protein sequences from the complete genomes.



**Fig. 3.** NJ tree of mitochondrial genomes based on log-correlation distance in the case  $K=12$  using the whole genome DNA sequences.

and primates are grouped together respectively. The non-eutherians [Marsupalia (Dvir and Mrob) and Monotremata (Oana)] are located at the root of all the mammals included in the study, which is similar to the results previously reported (Murphy *et al.* 2001b; Reyes *et al.* 2000; Stuart, Moffet and Baker 2002; Stuart, Moffet and Leader 2002). Interestingly, the elephant (Lafr) and aardvark (Oafe) group together as a branch since the elephant is a perissodactyl while the affinity of aardvark to other mammals is controversial. The two insectivores (Eur and Teur) failed to group together. The rabbit (Ocun) is found to be close to rodents. Although all rodents including the guinea pig (Cpor) are quite close to one another, they are not grouped as a branch. The monophyly of rodents (Reyes *et al.* 2000) is not well resolved by our method. Similarly, in the trees presented by Li *et al.* (2001) and Stuart, Moffet and Leader (2002), the guinea pig is not close to other rodents. The overall topology of the mammalian group of the tree in Fig. 1 is very similar to that derived from the SVD method (Stuart, Moffet

and Baker 2002; Stuart, Moffet and Leader 2002). In the remaining taxa, the overall topology in our tree is similar to that in the SVD tree (Stuart, Moffet and Leader 2002). The fish, reptiles and birds cluster as distinct groups as expected (**Fig. 1**). Within the birds, the falcon (Fper) groups with the storks (Cboy and Ccic); the redhead (Aame) groups with the chicken (Ggal); and *Rhea americana* (Rame) and *Struthio camelus* (Scam) group together. While the above three groupings are as expected, the interrelationships among the groups are not consistent with traditional view in which for example, *Rhea* should be a basal group in birds. The oscines *Covus* (cfu) and *Vidua* (Vcha) group together but their close relative *Smithornis sharpie* (Ssha) is not included in this group. Within the reptiles, the three turtles (Cmyd, Cpik and Psub) group together as a branch, but the close relationship of skink (Eegr) with the alligator (Amis) rather than snake (Dsem) is in contrast to the traditional view. The close relationship between the turtles and birds is also puzzling. In the cluster of fish, the chondrichthyes (cartilaginous fish) cluster as a group but osteichthyes (bony fish) are separated as two clades by the branch of chondrichthyes. The relationships among cartilaginous fish are the same as those in Stuart, Moffet and Leader (2002). The coelacanth (Lcha) and bichir (Porn) group together and constitute the basal branch of the fish cluster. The overall phylogeny of fish, including the relationship between cartilaginous fish and bony fish, is currently uncertain (Stuart, Moffet and Leader 2002). Yet the overall phylogeny of fish in our tree is different from that constructed by Rasmussen and Arnason (1999) and Stuart, Moffet and Leader (2002). The interrelationships among the mitochondrial genomes in our trees are roughly in agreement with the current understanding on vertebrate phylogeny.

Generally speaking the trees of **Figs. 2** and **3** are similar in topology to the tree shown in **Fig. 1**. In **Fig. 2**, *Alligator mississippiensis* (Amis) stays at the right place to be the closest species to birds in a group which is better than that in the tree in **Fig. 1**, *Falco peregrinus* (Fper) and *Danio rerio* (Drer) stay at wrong places, reptiles are separated by the branch of birds, and catartiodactyls are separated into two branches; the order of the branch of primates and the branch of non-eutherians should be switched. These are worse than the tree of **Fig. 1**. In **Fig. 3**, *Alligator mississippiensis* (Amis) stays at the right place to be the closest species to birds in the group of reptiles, while the bony fish are not separated by the branch of cartilaginous fish; this is better than the tree of **Fig. 1**. But *Falco peregrinus* (Fper) and *Sus scrofa* (Sscr) stay at the wrong places, which is worse than the tree of **Fig. 1**. Based on the above comparison, the tree of **Fig. 2** is a little bit worse than the trees of **Figs 1** and **3**, while the tree of **Fig. 3** is a little bit worse than that of **Fig. 1**. Hence the dynamical language model with correlation distance method is better than the other two methods for the data set selected.

We also tried replacing the Kullback-Leibler divergence distance by the correlation distance in method 2. The same result was obtained as shown in **Fig. 2**. In order to compare the dynamical language model approach with the Markov model approach proposed by Qi, Wang and Hao (2004), we generated all the trees of the same values of *K* based on these three kinds of data using the Markov model approach. But no tree generated by the Markov model approach can separate the major groups of mammals, namely fish, birds and reptiles, clearly. So for the mitochondrial genomes data set analyzed here, the methods introduced in this paper work better than the Markov model approach from the biological point of view. Our simple distance analyses on the complete mitochondrial genomes have yielded trees that are in roughly agreement with current knowledge on the phylogenetic relationships in different groups of vertebrates as elucidated previously by traditional analyses of the mitochondrial genomes and other molecular/ultrastructural approaches. Our approach circumvents the ambiguity in the selection of genes from complete genomes for phylogenetic reconstruction, and is also faster than the traditional approaches of phylogenetic analysis, particularly when dealing with a large number of genomes. Moreover, since multiple sequence alignment is not necessary, the intrinsic problems associated with this complex procedure can be avoided. Comparing with the method proposed in Li *et al.* (2001), our methods are more direct and faster, and the results are better from the biological point of view.

## Acknowledgments

Financial support was provided by the Chinese National Natural Science Foundation (grant no. 30570426), Fok Ying Tung Education Foundation (grant no. 101004) and the Youth Foundation of the Education Department of Hunan Province, China (grant no. 05B007) (Z.-G. Yu), Australian Research Council (grant no. DP0559807) (V.V. Anh), and the Research Grants Council of the Hong Kong Special Administrative Region, China (project no. CUHK4419/04M). The use of the facilities of Queensland University of Technology and Xiangtan University are gratefully acknowledged.

## References

- [1] U. Arnason, A. Gullberg, S. Gretarsdottir, B. Ursing and A. Janke, The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. **J. Mol. Evol.** Vol.50, 2000, pp 569–578.
- [2] R.L. Charlebois, R.G. Beiko and M. A. Ragan, Branching out. **Nature**, vol. 421, 2003, pp 217–217.
- [3] K.H. Chu, C.P. Li. and J. Qi, Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment. **Bioinformatics**, Vol. 22(14), 2006, pp 1690–1710.

- [4] K.H. Chu, J. Qi, Z.-G Yu and V. Anh, Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes. **Mol. Biol. Evol.**, vol.21, 2004, pp 200-206.
- [5] T. M. Cover and J. A. Thomas, **Elements of Information Theory**. New York: Wiley, 1991.
- [6] S. V. Edwards, B. Fertil, A. Giron, and P. Deschavanne J., A genomic schism in birds revealed by phylogenetic analysis of DNA strings. **Syst. Biol.**, Vol. 51, 2002, pp 599-613.
- [7] J.A. Eisen, and C.M. Fraser, Phylogenomics: intersection of evolution and genomics. **Science**, Vol. 300, 2003, pp 1706-1707.
- [8] J. Felsenstein, **PHYLIP** (phylogeny Inference package) version 3.5c. Distributed by the author at <http://evolution.genetics.washington.edu/phylip.html>, 1993.
- [9] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny. **Bioinformatics**, Vol. 17, 2001, pp 149-154.
- [10] S.K. Mouchaty, A. Gullberg, A. Janke, and U. Arnason, The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences. **Mol. Biol. Evol.** Vol. 17, 2000, pp 60–67.
- [11] W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, and S.J. O'Brien, Molecular phylogenetics and the origins of placental mammals. **Nature**, Vol. 409, 2001a, pp 614–618.
- [12] W.J. Murphy, E. Eizirik, S.J. O'Brien, et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. **Science**, Vol. 294, 2001b, pp 2348–2350.
- [13] M. Nikaido, M.M. Harad, Y. Cao, M. Hasegawa, and N. Okada, Monophyletic origin of the order chiroptera and its phylogenetic position among mammalia, as inferred from the complete sequence of the mitochondrial DNA of a japanese megabat, the ryukyu flying fox *Pteropus dasymallus*. **J. Mol. Evol.** Vol. 51, 2000, pp 318–328.
- [14] D.D. Pollock, J.A. Eisen, N.A. Doggett, and M.P. Cummings, A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. **Mol. Biol. Evol.** Vol. 17, 2000, pp 1776–1788.
- [15] J. Qi, H. Luo, and B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes. **Nucleic Acids Research**, Vol. 32, 2004, pp W45-W47.
- [16] J. Qi, B. Wang, and B. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. **J. Mol. Evol.**, Vol. 58, 2004, pp 1-11.
- [17] A.S. Rasmussen, and U. Arnason, Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree. **Proc. Natl. Acad. Sci. USA**, Vol. 965, 1999, pp 2177–2182.
- [18] A. Reyes, C. Gissi, G. Pesole, F.M. Catzeflis, and C. Saccone, Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. **Mol. Biol. Evol.**, Vol.17, 2000, pp 979–983.
- [19] A. Reyes, G. Pesole, and C. Saccone, Complete mitochondrial DNA sequence of the fat dormouse, *Glis glis*: further evidence of rodent parahyly. **Mol. Biol. Evol.** Vol. 15, 1998, pp 499-505.
- [20] N. Saitou, and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Mol. Biol. Evol.** Vol. 4, 1987, pp 406-425.
- [21] G.W. Stuart, K. Moffet, and S. Baker, Integrated gene species phylogenies from unaligned whole genome protein sequences. **Bioinformatics**, Vol. 18, 2002, pp 100-108.
- [22] G.W. Stuart, K. Moffet, and J. J. Leader, A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. **Mol. Biol. Evol.** Vol.19, 2002, pp 554-562.
- [23] O. Weiss, M.A. Jimenez and H. Herzog, Information content of protein sequences. **J. Theor. Biol.** Vol. 206, 2000, pp 379-386.
- [24] C.R. Woese, O. Kandler and M.L. Wheelis, Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya, **Proc. Natl. Acad. Sci. USA**, Vol. 87, 1990, pp 4576-4579.
- [25] X. Xu, A. Janke, and U. Arnason, The complete mitochondrial DNA sequence of the greater indian rhinoceros, *Rhinoceros unicornis*, and the phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla. **Mol. Biol. Evol.** Vol. 13, 1996, pp 1167–1173.
- [26] Z.G. Yu, V.V. Anh and K.S. Lau, Multifractal and correlation analysis of protein sequences from complete genome, **Phys. Rev. E**, vol 68], 2003, pp 021913.
- [27] Z.G. Yu, V.V. Anh and K.S. Lau, Chaos game representation, and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model, **J. Theor. Biol.**, vol 226(3), 2004, pp 341-348.
- [28] Z.G. Yu, V. Anh, K.S. Lau and K. H. Chu, The genomic tree of living organisms based on a fractal model. **Phys. Lett. A**, Vol. 317, 2003, pp 293-302.
- [29] Z.G. Yu and P. Jiang, Distance, correlation and mutual information among portraits of organisms based on complete genomes. **Phys. Lett. A**, Vol. 286, 2001, pp 34-46.
- [30] Z.G. Yu, L.Q. Zhou, V. V. Anh, K.H. Chu, S.C. Long and J.Q. Deng, Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment, **J. Mol. Evol.** Vol. 60, 2005, 538-545.